# SECRETED PROTEINS OF *MYCOBACTERIUM TUBERCULOSIS* AND THEIR USE AS VACCINES AND DIAGNOSTIC REAGENTS

## Background of the Invention

The invention is in the field of tuberculosis and, specifically, reagents useful for generating immune responses to *Mycobacterium tuberculosis* and for diagnosing infection and disease in a subject that has been exposed to *M. tuberculosis*.

Tuberculosis infection continues to be a world-wide health problem. This situation has recently been greatly exacerbated by the emergence of multi-drug resistant strains of *M. tuberculosis* and the international AIDS epidemic. It has thus become increasingly important that effective vaccines against and reliable diagnostic reagents for *M. tuberculosis* be produced.

U.S. application no. 08/796,792 is incorporated herein by reference in it entirety.

## Summary of the Invention

The invention is based on the discovery of a novel group of open reading frames (ORFs) encoding polypeptides that are secreted by *M. tuberculosis*. The invention features these polypeptides, functional segments thereof, DNA molecules encoding either the polypeptides or the functional segments, vectors containing the DNA molecules, cells transformed by the vectors, compositions containing one or more of any of the above polypeptides, functional segments, or DNA molecules, and a variety of diagnostic, therapeutic, and prophylactic (vaccine) methodologies utilizing the foregoing.

Specifically, the invention features an isolated DNA molecule containing a DNA sequence encoding a polypeptide

with a first amino acid sequence that can be the amino acid sequence of the polypeptide MTSP1, MTSP2, MTSP3, MTSP4, MTSP5, MTSP6, MTSP7, MTSP8, MTSP9, MTSP10, MTSP11, MTSP12, MTSP13, MTSP14, MTSP15, MTSP16, MTSP17, MTSP18, MTSP19, MTSP20, MTSP21, MTSP22, MTSP23, MTSP24, MTSP25, MTSP26, MTSP27, MTSP28, MTSP29, MTSP30, MTSP31, MTSP32,MTSP33, MTSP34, MTSP35, MTSP36, MTSP37, MTSP38, MTSP39, MTSP40, MTSP41, MTSP42, MTSP43, MTSP44, MTSP45, MTSP46, or MTSP47, as depicted in Fig. 1, or a second amino acid sequence identical to the first amino acid sequence with conservative substitutions; the polypeptide has *Mycobacterium tuberculosis* specific antigenic and immunogenic properties. Also included in the invention is an isolated portion of the above DNA molecule. The portion of the DNA molecule encodes a segment of the polypeptide shorter than the full-length polypeptide, and the segment has *Mycobacterium tuberculosis* specific antigenic and immunogenic properties. Other embodiments of the invention are vectors containing the above DNA molecules and transcriptional and translational regulatory sequences operationally linked to the DNA sequence, the regulatory sequences allow for the expression of the polypeptide or functional segment encoded by the DNA sequence in a cell. The invention encompasses cells (e.g., eukaryotic and prokaryotic cells) transformed with the above vectors.

The invention encompasses compositions containing any of the above vectors and a pharmaceutically acceptable diluent or filler. Other compositions to be used as DNA vaccines can contain at least two (e.g., three, four, five, six, seven, eight, nine, then, twelve, fifteen or twenty) DNA sequences, each encoding a polypeptide of the *Mycobacterium tuberculosis* complex or a functional segment thereof, with the DNA sequences being operationally linked to transcriptional and

translational regulatory sequences which allow for expression of each of the polypeptides in a cell of a vertebrate. In such compositions, at least one of the DNA sequences contains the sequence of the above DNA molecules of the invention.

The invention also features an isolated polypeptide with a first amino acid sequence that can be the sequence of the polypeptide MTSP1, MTSP2, MTSP3, MTSP4, MTSP5, MTSP6, MTSP7, MTSP8, MTSP9, MTSP10, MTSP11, MTSP12, MTSP13, MTSP14, MTSP15, MTSP16, MTSP17, MTSP18, MTSP19, MTSP20, MTSP21, MTSP22, MTSP23, MTSP24, MTSP25, MTSP26, MTSP27, MTSP28, MTSP29, MTSP30, MTSP31, MTSP32, MTSP33, MTSP34, MTSP35, MTSP36, MTSP37, MTSP38, MTSP39, MTSP40, MTSP41, MTSP42, MTSP43, MTSP44, MTSP45, MTSP46, or MTSP47, as depicted in Fig. 1, or a second amino acid sequence identical to the first amino acid sequence with conservative substitutions. The polypeptide has *Mycobacterium tuberculosis* specific antigenic and immunogenic properties. Also included in the invention is an isolated segment of this polypeptide, the segment being shorter than the full-length polypeptide and having *Mycobacterium tuberculosis* specific antigenic and immunogenic properties. Other embodiments are compositions containing the polypeptide, or functional segment, and a pharmaceutically acceptable diluent or filler. Compositions of the invention can also contain at least two (e.g., three four, five, six, seven, eight, nine, ten, twelve, fifteen, or twenty) polypeptides of the *Mycobacterium tuberculosis* complex, or functional segments thereof, with at least one of the at least two polypeptides having the sequence of one of the above described polypeptides of the invention.

The invention also features methods of diagnosis. One embodiment is a method involving: (a) administration of one of the above polypeptide compositions to a subject suspected

of having or being susceptible to *Mycobacterium tuberculosis* infection; and (b) detecting an immune response in the subject to the composition, as an indication that the subject has or is susceptible to *Mycobacterium tuberculosis* infection. Another embodiment is a method that involves: (a) providing a population of cells containing CD4 T lymphocytes from a subject; (b) providing a population of cells containing antigen presenting cells (APC) expressing a major histocompatibility complex (MHC) class II molecule expressed by the subject; (c) contacting the CD4 lymphocytes of (a) with the APC of (b) in the presence of one or more of the polypeptides, functional segments, and or polypeptide compositions of the invention; and (d) determining the ability of the CD4 lymphocytes to respond to the polypeptide, as an indication that the subject has or is susceptible to *Mycobacterium tuberculosis* infection. Another diagnostic method of the invention involves: (a) contacting a polypeptide, a functional segment, or a polypeptide/functional segment composition of the invention with a bodily fluid of a subject; (b) detecting the presence of binding of antibody to the polypeptide, functional segment, or polypeptide/functional segment composition, as an indication that the subject has or is susceptible to *Mycobacterium tuberculosis* infection.

Also encompassed by the invention are methods of vaccination. These methods involve administration of any of the above polypeptides, functional segments, or DNA compositions to a subject. The compositions can be administered alone or with one or more of the other compositions.

As used herein, an "isolated DNA molecule" is a DNA which is one or both of: not immediately contiguous with one

or both of the coding sequences with which it is immediately contiguous (i.e., one at the 5' end and one at the 3' end) in the naturally-occurring genome of the organism from which the DNA is derived; or which is substantially free of DNA sequence with which it occurs in the organism from which the DNA is derived. The term includes, for example, a recombinant DNA which incorporated into a vector, e.g., into an autonomously replicating plasmid or virus, or into the genomic DNA of a prokaryote or eukaryote, or which exists as a separate molecule (e.g., a cDNA or a genomic fragment produced by PCR or restriction endonuclease treatment) independent of other DNA sequences. Isolated DNA also includes a recombinant DNA which is part of a hybrid DNA encoding additional *M. tuberculosis* polypeptide sequences.

"DNA molecules" include cDNA, genomic DNA, and synthetic (e.g., chemically synthesized) DNA. Where single-stranded, the DNA molecule may be a sense strand or an antisense strand.

An "isolated polypeptide" of the invention is a polypeptide which either has no naturally-occurring counterpart, or has been separated or purified from components which naturally accompany it, e.g., in *M. tuberculosis* bacteria. Typically, the polypeptide is considered "isolated" when it is at least 70%, by dry weight, free from the proteins and naturally-occurring organic molecules with which it is naturally associated. Preferably, a preparation of a polypeptide of the invention is at least 80%, more preferably at least 90%, and most preferably at least 99%, by dry weight, the peptide of the invention. Since a polypeptide that is chemically synthesized is, by its nature, separated from the components that naturally accompany it, the synthetic polypeptide is "isolated."

An isolated polypeptide of the invention can be obtained, for example, by extraction from a natural source (e.g., *M. tuberculosis* bacteria); by expression of a recombinant nucleic acid encoding the polypeptide; or by chemical synthesis. A polypeptide that is produced in a cellular system different from the source from which it naturally originates is "isolated," because it will be separated from components which naturally accompany it. The extent of isolation or purity can be measured by any appropriate method, e.g., column chromatography, polyacrylamide gel electrophoresis, or HPLC analysis.

The polypeptides may contain a primary amino acid sequence that has been modified from those disclosed herein. Preferably these modifications consist of conservative amino acid substitutions. Conservative substitutions typically include substitutions within the following groups: glycine and alanine; valine, isoleucine, and leucine; aspartic acid and glutamic acid; asparagine and glutamine; serine and threonine; lysine and arginine; and phenylalanine and tyrosine.

The terms "protein" and "polypeptide" are used herein to describe any chain of amino acids, regardless of length or post-translational modification (for example, glycosylation or phosphorylation). Thus, the term "*Mycobacterium tuberculosis* polypeptide" includes full-length, naturally occurring *Mycobacterium tuberculosis* protein, as well a recombinantly or synthetically produced polypeptide that corresponds to a full-length naturally occurring *Mycobacterium tuberculosis* protein or to particular domains or portions of a naturally occurring protein. The term also encompasses a mature *Mycobacterium tuberculosis* polypeptide

which has an added amino-terminal methionine (useful for expression in prokaryotic cells).

As used herein, "immunogenic" means capable of activating a primary or memory immune response. Immune responses include responses of CD4+ and CD8+ T lymphocytes and B-lymphocytes. In the case of T lymphocytes, such responses can be proliferative, and/or cytokine (e.g., interleukin(IL)-2, IL-3, IL-4, IL-5, IL-6, IL-12, IL-13, IL-15, tumor necrosis factor-α (TNF-α), or interferon-γ (IFN-γ))-producing, or they can result in generation of cytotoxic T-lymphocytes (CTL). B-lymphocyte responses can be those resulting in antibody production by the responding B lymphocytes.

As used herein, "antigenic" means capable of being recognized by either antibody molecules or antigen-specific T cell receptors (TCR) on activated effector T cells (e.g., cytokine-producing T cells or CTL).

Thus, polypeptides that have "*Mycobacterium tuberculosis* specific antigenic properties" are polypeptides that: (a) can be recognized by and bind to antibodies elicited in response to *Mycobacterium tuberculosis* organisms or wild-type *Mycobacterium tuberculosis* molecules (e.g., polypeptides); or (b) contain subsequences which, subsequent to processing of the polypeptide by appropriate antigen presenting cells (APC) and bound to appropriate major histocompatibility complex (MHC) molecules, are recognized by and bind to TCR on effector T cells elicited in response to *Mycobacterium tuberculosis* organisms or wild-type *Mycobacterium tuberculosis* molecules (e.g., polypeptides).

As used herein, polypeptides that have "*Mycobacterium tuberculosis* specific immunogenic properties" are polypeptides that: (a) can elicit the production of

antibodies that recognize and bind to *Mycobacterium tuberculosis* organisms or wild-type *Mycobacterium tuberculosis* molecules (e.g., polypeptides); or (b) contain subsequences which, subsequent to processing of the polypeptide by appropriate antigen presenting cells (APC) and bound to appropriate major histocompatibility complex (MHC) molecules on the surface of the APC, activate T cells with TCR that recognize and bind to peptide fragments derived by processing by APC of *Mycobacterium tuberculosis* organisms or wild-type *Mycobacterium tuberculosis* molecules (e.g., polypeptides) and bound to MHC molecules on the surface of the APC. The immune responses elicited in response to the immunogenic polypeptides are preferably protective. As used herein, "protective" means preventing establishment of an infection or onset of a disease or lessening the severity of a disease existing in a subject. "Preventing" can include delaying onset, as well as partiallly or completely blocking progress of the disease.

As used herein, a "functional segment of a *Mycobacterium tuberculosis* polypeptide" is a segment of the polypeptide that has *Mycobacterium tuberculosis* specific antigenic and immunogenic properties.

Where a polypeptide, functional segment of a polypeptide, or a mixture of polypeptides and/or functional segments have been administered (e.g., by intradermal injection) to a subject for the purpose of testing for a *M. tuberculosis* infection or susceptibility to such an infection, "detecting an immune response" means examining the subject for signs of a immunological reaction to the administered material, e.g., reddening or swelling of the skin at the site of an intradermal injection. Where the subject has antibodies to the administered material, the

response will generally be rapid, e.g., 1 minute to 24 hours. On the other hand, a memory or activated T cell reaction of pre-immunized T lymphocytes in the subject is generally slower, appearing only after 24 hours and being maximal at 24-96 hours.

As used herein, a "subject" can be a human subject or a non-human mammal such as a non-human primate, a horse, a bovine animal, a pig, a sheep, a goat, a dog, a cat, a rabbit, a guinea pig, a hamster, a rat, or a mouse.

Unless otherwise defined, all technical and scientific terms used herein have the same meaning as commonly understood by one of ordinary skill in the art to which this invention pertains. In case of conflict, the present document, including definitions, will control. Preferred methods and materials are described below, although methods and materials similar or equivalent to those described herein can be used in the practice or testing of the present invention. Unless otherwise indicated, these materials and methods are illustrative only and are not intended to be limiting. All publications, patent applications, patents and other references mentioned herein are illustrative only and not intended to be limiting.

Other features and advantages of the invention, e.g., methods of diagnosing or vaccinating against *M. tuberculosis* infection, will be apparent from the following description, from the drawings and from the claims.

## Brief Description of the Drawings

Figure 1 is a depiction of the amino acid sequences of *M. tuberculosis* polypeptides MTSP1-MTSP47.

Figure 2 is a depiction of the nucleotide sequences of the coding regions (mtsp1-mtsp47) encoding MTSP1-MTSP47.

Fig. 3A is a line graph showing the distribution of SPSCAN scores for the 3924 *M. tuberculosis* protein sequences obtained from the Sanger Center website.

Fig. 3B is a line graph showing the distribution of SignalP scores for the 3924 protein sequences obtained from the Sanger Center website.

Fig. 3C is a "dot plot" of SignalP scores versus SPSCAN scores for the individual 3924 protein sequences obtained from the Sanger Centre website.

Fig. 4 is an enlargement of Fig. 3C.

## Detailed Description

It is generally believed that proteins that are actively secreted by bacteria, especially intracellular bacteria (e.g., *Salmonella typhi* and *M. tuberculosis*), are effective as antigens that are capable of inducing protective immunity to the organism. A number of open reading frames (ORF), (i.e., DNA sequences that encode a protein) were predicted from the genomic sequence of *M. tuberculosis* [Cole et al. (1998) Nature 393:537-544]. The instant invention is based on the identification of a number of ORFs of this group that encode secreted polypeptides (see Example 1). The polypeptides encoded by the ORFs thus identified are designated *M. tuberculosis* Secreted Polypeptides (MTSP) and the DNA sequences encoding them are designated mtsp. Because they are secreted, we believe that the MTSP are both immunogenic and antigenic. The immune responses that they induce in subjects exposed to them are preferably also protective against *M. tuberculosis* infection in the subjects. The amino acid sequences of MTSP1-MTSP44 are shown in Fig. 1 and the nucleotide sequences of mtsp1-mtsp44 are shown in Fig. 2.

The invention encompasses: (a) isolated DNA molecules containing sequences (e.g., mtsp1-mtsp47) encoding polypeptides (e.g., MTSP1-MTSP47) secreted by *M. tuberculosis* and isolated portions of such DNA molecules that encode polypeptide segments having antigenic and immunogenic properties (i.e., functional segments); (b) the secreted polypeptides themselves (e.g., MTSP1-MTSP47) and functional segments of them; (c) antibodies (including antigen binding fragments, e.g., $F(ab')_2$, Fab, Fv, and single chain Fv fragments of such antibodies) that bind to the MTSP1-MTSP47 polypeptides and functional segments; (d) nucleic acid molecules (e.g., vectors) containing and capable of expressing one or more of the DNA molecules containing the mtsp1-mtsp47 sequences and portions of DNA molecules; (e) cells (e.g., bacterial, yeast, insect, or mammalian cells) transformed by such vectors; (f) compositions containing vectors encoding one or more *M. tuberculosis* polypeptides (or functional segments) including both the MTSP1-MTSP47 polypeptides (or functional segments thereof) and previously described *M. tuberculosis* polypeptides such as ESAT-6, 14 kDa antigen, MPT63, 19 kDa antigen, MPT64, MPT51, MTC28, 38 kDa antigen, 45/47 kDa antigen, MPB70, Ag85 complex, MPT53, and KatG (see also U.S. application no. 08/796,792); (g) compositions containing one or more *M. tuberculosis* polypeptides (or functional segments), including both the polypeptides of the invention and previously described *M. tuberculosis* polypeptides such as those described above; (h) compositions containing one or more of antibodies described in (c); (i) methods of diagnosis involving either (1) administration (e.g., intradermal injection) of the MTSP1-MTSP44 polypeptides of the invention, functional segments thereof, or mixtures of one more such polypeptides and/or

functional segments to a subject suspected of having or being susceptible to *M. tuberculosis* infection, (2) in vitro testing of lymphocytes from such a subject for responsiveness to the MTSP1-MTSP47 polypeptides, functional segments thereof, or the above mixtures, or (3) testing of a bodily fluid (e.g., blood, saliva, plasma, serum, urine, or semen or a lavage such as a bronchoalveolar lavage, a vaginal lavage, or lower gastrointestinal lavage) for antibodies to the MTSP1-MTSP47 polypeptides or functional segments thereof, or the above-described mixtures; (j) methods of vaccination involving administration to a subject of the compositions of either (f), (g), (h) or a combination of any two or even all 3 compositions.

With respect to diagnosis, purified *M. tuberculosis* proteins, functional segments of such proteins, or mixtures of proteins and/or the functional fragments have the advantage of discriminating infection by *M. tuberculosis* from infection by other bacteria, and in particular, non-pathogenic mycobacteria. Of particular benefit in such assays are proteins encoded by genes present in *M. tuberculosis*, and possibly other members of the M.tuberculosis complex (e.g., *M. tuberculosis*, M. bovis, M. microti, and M. africanum), but absent from the Bacille Calmette-Guerin (BCG) attenuated strain of M. bovis which has been commonly used for vaccination. Use of such proteins (e.g., the MTSP16 protein whose sequence is shown in Fig. 1) for diagnosis allows for discrimination between infection by *M. tuberculosis* and vaccination with BCG. Furthermore, compositions containing the *M. tuberculosis* proteins, functional segments of them, or mixtures of the proteins and/or the functional segments allows for improved quality control since "batch-to-batch" variability is greatly reduced

in comparison to complex mixtures such as purified protein derivative (PPD) of tuberculin.

Where vaccination is performed with nucleic acids both in vivo and ex vivo methods can be used. In vivo methods involve administration of the nucleic acids themselves to the subject and ex vivo methods involve obtaining cells (e.g., bone marrow cells or fibroblasts) from the subject, transducing the cells with the nucleic acids, preferably selecting or enriching for successfully transduced cells, and administering the transduced cells to the subject. Alternatively, the cells that are transduced and administered to the subject can be derived from another subject. Methods of vaccination and diagnosis are described in greater detail in U.S. application no. 08/796,792 which is incorporated herein by reference in its entirety.

The following example is meant to illustrate, not limit the invention.

**Example 1. Computer Aided Identification of *M. tuberculo sis* Secreted Proteins**

Software.

The software used to manipulate and analyze protein sequences was available from public web servers or was part of the Genetics Computer Group (GCG) package [Wisconsin Package Version 9.1, Genetics Computer Group (GCG), Madison, Wisc.]. Customized C-Shell scripts were used to automate some of the tasks or to extract selected information from the output of some of the programs. Signal peptides were predicted with SPSCAN, which is part of the GCG package, and SignalP, a program originating from the Center for Biological Sequence Analysis at the Technical University of Denmark, Lyngby, Denmark and currently available on the Internet at http://www.cbs.dtu.dk/services/SignalP. Putative

transmembrane segments were identified with the program TMpred and prokaryotic membrane lipoprotein lipid attachment sites with the program PrositeScan, both programs originating from the Bioinformatics Group at the Swiss Institute for Experimental Cancer Research in Epalinges, Switzerland, and currently available on the Internet at http://www.isrec.isb-sib.ch/software/TMPRED_form.html and http://www.isrec.isb-sib.ch/software/PSTSCAN_form.html, respectively. Protein similarity and relatedness was established with GAP and PILEUP, both in the GCG package, Blast originating from the National Center for Biotechnology Information of the National Institutes for Health, Bethesda, MD and currently available on the Internet at http://www.ncbi.nlm.nih.gov/BLAST/, and AllAll originating from the Swiss Institute of Technology, Zurich, Switzerland, and currently available on the Internet at http:cbrg.inf.ethz.ch/subsection3_1_1.html.

Prediction of *M. tuberculosis* proteins with signal peptides

The amino acid sequences of the 3924 proteins predicted by the analysis of the *M. tuberculosis* genomic sequence have been made available by the Sanger Centre, Cambridge, England, and were downloaded from the current Sanger Center website [http://www.sanger.ac.uk/Projects/M_tuberculosis/]. Segments containing the first 70 amino acids of each predicted protein were analyzed by a system of our own design utilizing two different computer programs (SPSCAN and SignalP) designed to predict the occurrence of signal peptides. We concluded that combining the output from the two programs would increase the reliability of the selection. Both programs can detect signal peptides in polypeptides from eukaryotic and prokaryotic organisms, including gram-positive and gram-negative bacteria. To analyze the *M. tuberculosis* proteins the gram-positive mode was used. We performed an analysis

with SPSCAN allowing only one prediction per protein, setting the minimum score threshold at -10, both in the standard and the adjusted modes. In the adjusted mode, signal peptides longer than a certain threshold value are penalized. We found that the correlation between the scores obtained with SPSCAN in the standard and adjusted modes increased with the value of the score, i.e., signal peptides that received high scores in standard mode also had high scores in the adjusted mode. We determined to use only the adjusted mode in subsequent steps.

To define cutoff values for the scores obtained with SPSCAN (in adjusted mode) and SignalP we took into account the following factors: (a) SignalP scores above 0.34 are generally considered significant; (b) the analysis of Haemophilus influenzae genome with SignalP yielded the prediction that about 10% of the encoded proteins contain a signal peptide; and (c) the average scores of thirteen known secreted or membrane-associated *M. tuberculosis* antigens was 9.11 (standard deviation (SD)=1.8) and 0.55 (SD=0.15), as calculated as above utilizing SPSCAN and SignalP, respectively (Table 1).

Of the 3924 *M. tuberculosis* protein sequences downloaded from the Sanger Centre website, about 10% of the sequences had SPSCAN scores equal or higher than 8 (Fig. 3A) and about 10% of the sequences had SignalP scores equal or higher than 0.4 (Fig 3B). We tentatively adopted these score values as "cutoffs" and we used the cutoffs to construct a list of proteins that were likely to be either secreted or exposed at the bacterial cell surface. This list included those proteins with SPSCAN scores higher than 8 and SignalP scores higher than 0.4. We refer to this group of proteins (208

entries, about 5% of the proteome) as the "Top208" group (Fig. 3C and Fig. 4).

Table 1. SPCAN and SignalP Scores of Known Secreted or Membrane Associated *M. tuberculosis* Polypeptide Antigens

| Polypeptide Antigens | Alternative Names | SPSCAN Score | SignalP Score |
|---|---|---|---|
| 19 kDa | | 5.9 | 0.331 |
| 38 kDa | PhoS, Ag78, antigen 5 | 6.3 | 0.505 |
| 45/47 kDa | | 11.2 | 0.627 |
| MPT44 | Ag85A, P32, FbpA | 9.2 | 0.425 |
| MPT45 | Ag85C, FbpC | 10.1 | 0.496 |
| MPT51 | | 11 | 0.758 |
| MPT53 | | 9.4 | 0.581 |
| MPT59 | Ag85B, á antigen, Ag 6, FbpB | 9.7 | 0.629 |
| MPT63 | | 8 | 0.57 |
| MPT64 | | 10.2 | 0.83 |
| MPT70 | | 9 | 0.459 |
| MPT83 | | 7.1 | 0.298 |
| MTC28 | | 11.4 | 0.7 |

Prediction of *M. tuberculosis* secreted proteins

A signal peptide may target a protein to the membrane but does not define a secreted protein, because additional transmembrane segments within the mature protein molecule can be present. In addition, lipoproteins are also targeted to the membrane by a signal peptide, but are not all secreted since cleavage of the signal peptide is coupled with the attachment of an acyl glycerol group that anchors the protein to the membrane. In light of these considerations and the fact that SignalP is not designed to differentiate lipoprotein signal peptides from secretory signal peptides, we believe that the Top208 group contains lipoproteins and proteins with multiple transmembrane segments, in addition to secreted proteins.

The number of putative transmembrane segments and the presence of lipoprotein lipid attachment sites were assessed by analyzing the Top208 proteins with TMpred and PrositeScan.

TMpred identifies putative transmembrane segments by comparing a query amino acid sequence with a database of amino acid sequences of experimentally defined transmembrane segments. Scores higher than 500 are considered significant. PrositeScan compares query amino acid sequences against the Prosite database of protein motifs. The prokaryotic lipoprotein lipid attachment site motif is entry number PS00013. Our methodology identified a class of secreted proteins (the "Top208-TM1" group that included MTSP1-MTSP44) which were characterized by a single transmembrane segment (with score higher than 500) in the position predicted for the signal peptide and in which no lipoprotein motifs were identified. Other proteins had additional transmembrane segments with scores higher than 500, had lipoprotein motifs, or were excluded from the analysis because they belonged to the PE/PPE/PGRS families of proteins [Cole et al., 1998] and their biased amino acid composition made it difficult to obtain reliable results with SPSCAN, SignalP, or TMpred. A summary of the characteristics of the proteins we assigned to the Top208-TM1 group is presented in Table 2 and data regarding proteins MTSP1-MTSP47 are presented in Table 3. The amino acid sequences of the proteins are listed in Fig. 1 and the nucleotide sequences of ORF encoding them (mtsp1-mtsp47) are listed in Fig. 2.

Table 2. Features defining the *M. tuberculosis* proteins included in the Top208-TM1 group.

1. A signal peptide with score higher than 0.4 was predicted with SignalP in the first 70 amino acids.
2. A signal peptide with score higher than 8 was predicted with SPSCAN in the first 70 amino acids.
3. A single transmembrane segment, with a score greater than 500 and coinciding approximately with the putative signal peptide, was predicted by TMpred.
4. No lipoprotein lipid attachment sites were identified with PrositeScan.

Table 3. Proteins included in the Top208-TM1 group.

| Protein | No. of Amino Acids | SPSCAN Score | SPSCAN Sequence | SignalP Score | SignalP Sequence |
|---|---|---|---|---|---|
| MTSP20 | 130 | 12.4 | 1-32 | 0.672 | 1-32 |
| MTSP21 | 109 | 8.4 | 1-22 | 0.631 | 1-22 |
| MTSP23 | 114 | 10.2 | 1-34 | 0.592 | 1-34 |
| MTSP16 | 126 | 9.2 | 1-28 | 0.557 | 1-36 |
| MTSP24 | 125 | 11.4 | 1-35 | 0.73 | 1-35 |
| MTSP14 | 144 | 8.9 | 1-34 | 0.584 | 1-34 |
| MTSP13 | 157 | 10 | 1-32 | 0.753 | 1-32 |
| MTSP22 | 124 | 8.6 | 1-30 | 0.592 | 1-30 |
| MTSP25 | 155 | 9.5 | 35-49 | 0.842 | 1-49 |
| MTSP27 | 233 | 13.8 | 1-29 | 0.787 | 1-29 |
| MTSP11 | 233 | 10.9 | 1-32 | 0.779 | 1-32 |
| MTSP26 | 382 | 8.3 | 1-34 | 0.721 | 1-34 |
| MTSP12 | 214 | 12.6 | 1-28 | 0.71 | 1-28 |
| MTSP8 | 158 | 9.1 | 1-33 | 0.695 | 1-30 |
| MTSP10 | 155 | 8.8 | 15-45 | 0.669 | 1-45 |
| MTSP28 | 295 | 14.8 | 1-31 | 0.667 | 1-31 |
| MTSP9 | 241 | 10 | 1-22 | 0.635 | 1-22 |
| MTSP29 | 380 | 12.4 | 1-27 | 0.621 | 1-27 |
| MTSP2 | 111 | 10.6 | 1-28 | 0.579 | 1-28 |
| MTSP4 | 177 | 8.7 | 1-25 | 0.578 | 1-24 |
| MTSP17 | 219 | 8.9 | 1-29 | 0.543 | 1-29 |
| MTSP3 | 282 | 11.5 | 1-32 | 0.538 | 1-32 |
| MTSP18 | 220 | 8.8 | 38-68 | 0.537 | 1-68 |
| MTSP6 | 219 | 8.4 | 1-34 | 0.537 | 1-34 |
| MTSP7 | 136 | 11.7 | 1-24 | 0.53 | 1-24 |
| MTSP31 | 457 | 9.1 | 1-18 | 0.494 | 1-25 |
| MTSP30 | 286 | 8.3 | 15-37 | 0.469 | 1-37 |
| MTSP1 | 104 | 8.2 | 1-28 | 0.466 | 1-28 |
| MTSP15 | 134 | 10 | 1-21 | 0.458 | 1-56 |
| MTSP32 | 449 | 8.8 | 1-23 | 0.444 | 1-23 |
| MTSP19 | 169 | 10.5 | 28-53 | 0.438 | 1-53 |
| MTSP5 | 568 | 9.9 | 1-31 | 0.432 | 1-31 |
| MTSP33 | 113 | 11.9 | 1-25 | 0.873 | 1-25 |
| MTSP41 | 112 | 12 | 1-33 | 0.663 | 1-3 |
| MTSP38 | 173 | 10.5 | 1-28 | 0.697 | 1-28 |
| MTSP35 | 408 | 8.8 | 1-33 | 0.616 | 1-33 |
| MTSP34 | 149 | 13.7 | 1-23 | 0.888 | 1-23 |
| MTSP36 | 168 | 11.3 | 1-28 | 0.824 | 1-27 |
| MTSP42 | 521 | 8.4 | 1-34 | 0.679 | 1-34 |
| MTSP44 | 149 | 11 | 1-30 | 0.661 | 1-30 |
| MTSP37 | 228 | 9.4 | 1-23 | 0.598 | 1-23 |
| MTSP40 | 231 | 9.2 | 1-30 | 0.55 | 1-30 |
| MTSP43 | 137 | 8.2 | 1-36 | 0.485 | 1-37 |
| MTSP39 | 509 | 8.6 | 1-35 | 0.413 | 1-38 |
| MTSP45 | 145 | 8.4 | 1-46 | 0.412 | 1-62 |
| MTSP46 | 143 | 8.5 | 1-27 | 0.555 | 1-66 |
| MTSP47 | 171 | 8.3 | 1-35 | 0.424 | 1-30 |

SPSCAN sequence and SignalP sequence show the sequence, in terms of amino acid residue numbers, included in the signal peptide predicted by SPSCAN and SignalP, respectively.

Table 4. Presence mtsp coding regions in various strains of *Mycobacterium tuberculosis*.

| Coding Region | *M. tuberculosis* | *M. bovis BCG* | *M. bovis* | *M. kansasii* | *M. africanum* | *M. scrofulaceum* | *M. fortuitum* | *M. marinum* | *M. malmoense* | *M. avium* | *M. gastri* | *M. chelonae* | *M. ulcerans* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MTSP6 | + | + | + | + | + | + | + | + | + | + | + | + | + |
| MTSP28 | + | + | + | + | + | + | + | + | + | + | + | + | + |
| MTSP44 | + | + | + | + | + | + | + | + | + | + | + | + | +/- |
| MTSP34 | + | + | + | +/- | + | + | + | + | + | + | + | + | + |
| MTSP39 | + | + | + | + | + | + | + | + | + | + | + | + | + |
| MTSP1 | + | + | + | + | + | + | + | + | + | + | + | + | + |
| MTSP15 | + | + | + | + | + | + | + | + | + | + | + | + | + |
| MTSP35 | + | + | + | + | + | + | + | + | + | + | + | + | + |
| MTSP5 | + | + | + | + | + | + | + | + | + | + | + | + | + |
| MTSP46 | + | + | + | + | + | - | - | + | - | - | + | - | - |
| MTSP11 | + | + | + | + | + | - | - | - | + | - | + | - | - |
| MTSP24 | + | + | + | + | + | + | - | + | + | - | + | + | + |
| MTSP23 | + | + | + | + | + | + | - | + | + | + | + | - | - |
| MTSP41 | + | + | + | + | + | - | - | - | + | - | + | - | - |
| MTSP22 | + | + | + | - | + | - | + | - | - | - | - | - | - |
| MTSP26 | + | + | + | + | + | + | + | + | + | + | + | + | + |
| MTSP40 | + | + | + | + | + | + | + | + | + | + | + | + | + |
| MTSP13 | + | + | + | + | + | + | + | + | + | + | - | - | - |
| MTSP16 | + | - | - | - | + | - | - | - | - | - | - | - | - |
| MTSP42 | + | + | + | + | + | + | - | + | + | - | + | + | + |
| MTSP36 | + | + | + | + | + | | | | | | | | |
| MTSP47 | + | + | + | - | + | - | - | - | - | - | - | - | - |
| MTSP38 | + | + | + | + | + | + | + | + | + | + | + | + | + |
| MTSP10 | + | + | + | + | + | + | + | + | + | + | + | + | + |
| MTSP37 | + | + | + | + | + | | | | | | | | |
| MTSP29 | + | + | + | + | + | + | + | + | + | + | + | +/- | +/- |
| MTSP31 | + | + | + | + | + | + | + | + | + | + | + | + | + |
| MTSP32 | + | + | + | + | + | + | + | + | + | + | + | + | + |
| MTSP30 | + | + | + | + | + | + | + | + | + | + | +/- | +/- | + |
| MTSP3 | + | + | + | + | + | + | + | + | + | + | + | + | + |
| MTSP20 | + | + | + | - | + | | | | | | | | |
| MTSP4 | + | + | + | + | + | + | + | + | + | + | + | + | + |
| MTSP27 | + | + | + | + | + | + | + | + | + | + | + | +/- | +/- |

The inventors have found, by standard DNA hybridization. Southern blotting techniques using the indicated coding regions as probes and DNA isolated from the indicated strains of *Mycobacteria*, that some of the coding regions are specific for the *M. tuberculosis* complex. (Table 4)

Although the invention has been described with reference to the presently preferred embodiment, it should be understood that various modifications can be made without departing from the spirit of the invention. Accordingly, the invention is limited only by the following claims.